(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号
特開2001－272991
（Ｐ２００１－２７２９９１Ａ）
(43)公開日　平成13年10月５日(2001.10.5)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード（参考）
G10L 13/06		G10L 5/04	F	5D015
13/00		3/00	R	5D045
13/08			H	9A001
15/00			551 Q	
15/28			571 T	

審査請求　未請求　請求項の数10　ＯＬ　（全11頁）　最終頁に続く

(21)出願番号　特願2000－84948(Ｐ2000－84948)

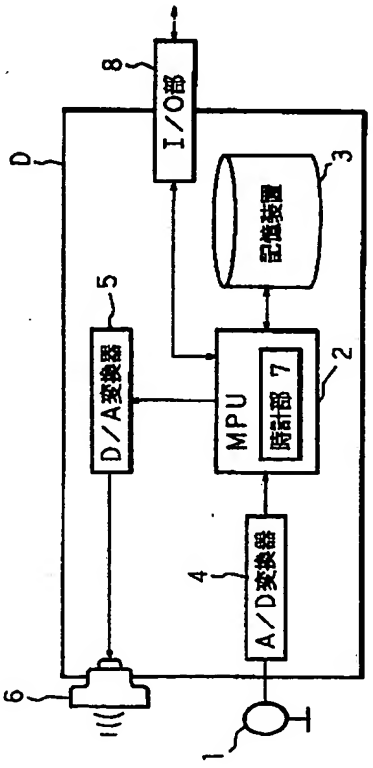
(22)出願日　平成12年３月24日(2000.3.24)

(71)出願人　000001889
三洋電機株式会社
大阪府守口市京阪本通２丁目５番５号
(72)発明者　橋本　誠
大阪府守口市京阪本通２丁目５番５号　三洋電機株式会社内
(74)代理人　100078868
弁理士　河野　登夫

Ｆターム(参考)　5D015 AA04 CC03 CC13 CC14 FF00
KK02 KK04 LL00
5D045 AA07 AB30
9A001 HH17 HH18 HZ19 JJ77 KK56

(54)【発明の名称】音声対話方法及び音声対話装置

(57)【要約】
【課題】　カーナビゲーション装置やテレビゲームなど音声対話機能搭載機器が登場しているが、音声対話の進行状況に応じて出力音声の声質や口調を変えるものはなく、同じ声質及び口調の合成音声が出力されるため、ユーザーが機械操作に飽きるという問題があった。
【解決手段】　入力音声の速度又は抑揚等を分析することによって話者の感情をカテゴリー化し、そのカテゴリーに対応させて応答音声をもカテゴリー化して感情を持った応答をさせるようにしたので、単調さが無くなり、あたかも人間と会話している様なユーザーフレンドリーな音声対話システムの構築が可能となる。さらに、入力頻度の高いキーワードに対応する応答単語ついては、入力頻度の低いキーワードに対応する応答単語よりも、強調して応答音声を出力させるようにしたので、話者は応答音声中重要なポイントを的確に聞くことができる。



1

【特許請求の範囲】

【請求項１】　入力された音声に対して応答文を作成して音声出力する音声対話方法において、
入力音声を音声信号に変換するステップと、
該音声信号を音声特徴情報に変換するステップと、
該音声特徴情報に基づいて所定のカテゴリ群から入力音声のカテゴリを決定するステップと、
前記入力音声のカテゴリに応じて所定の応答カテゴリを決定するステップと、
該応答カテゴリに応じて出力応答文の音声特徴情報を決定する特徴情報決定ステップと、
該特徴情報決定ステップにより決定した音声特徴情報に基づいて出力応答文の音声を合成するステップとを備えることを特徴とする音声対話方法。

【請求項２】　カテゴリを決定した回数をカテゴリ毎に計数するステップを更に備え、
前記特徴情報決定ステップは、前記応答カテゴリ及び前記ステップで計数した回数に応じて決定することを特徴とする請求項１に記載の音声対話方法。

【請求項３】　カテゴリを決定した時刻をカテゴリ毎に記憶するステップを更に備え、
前記特徴決定ステップは、前記応答カテゴリ、前記ステップで計数した回数及び前記ステップで記憶した時刻に応じて決定することを特徴とする請求項２に記載の音声対話方法。

【請求項４】　入力された音声に対して応答文を作成して音声出力する音声対話方法において、
入力音声を音声信号に変換するステップと、
該音声信号を音声特徴情報に変換するステップと、
該音声特徴情報に基づいて所定のキーワード及び音素群から入力音声のキーワード及び音素を決定するステップと、
同一キーワードが入力された回数を計数するステップと、
入力されたキーワード及び音素に基づいて応答文を作成するステップと、
作成した応答文のうち前記キーワードに対応して作成した応答単語について、キーワード入力回数に応じて出力する応答文中の各応答単語の韻律又はスペクトルを決定するステップと、
前記決定した出力応答文中の応答単語の韻律又はスペクトルに基づいて出力応答文の音声を合成するステップとを備えることを特徴とする音声対話方法。

【請求項５】　入力された音声に対して応答文を作成して音声出力する音声対話方法において、
入力信号を音声信号に変換するステップと、
該音声信号を音声特徴情報に変換するステップと、
該音声特徴情報に基づいて所定のキーワード及び音素群から入力音声のキーワード及び音素を決定するステップと、

2

同一キーワードが入力された回数を計数するステップと、
入力されたキーワード及び音素に基づいて応答文を作成するステップと、
作成した応答文のうち前記キーワードに対応して作成した応答単語について、キーワード入力回数に応じて出力する応答文中の各応答単語の韻律又はスペクトルを決定するステップと、
前記音声特徴情報に基づいて所定のカテゴリ群から入力音声のカテゴリを決定するステップと、
カテゴリを決定した回数をカテゴリ毎に計数するステップと、
カテゴリを決定した時刻をカテゴリ毎に記憶するステップと、
前記入力音声のカテゴリに応じて所定の応答カテゴリを決定するステップと、
前記応答カテゴリ、カテゴリを決定した前記ステップで計数した回数及び前記ステップで記憶した時刻に応じて出力応答文の音声特徴情報を決定するステップと、
該決定した出力応答文の音声特徴情報及び前記決定した出力応答文中の応答単語の韻律又はスペクトルに基づいて出力応答文の音声を合成するステップとを備えることを特徴とする音声対話方法。

【請求項６】　入力された音声に対して応答文を作成して音声出力する音声対話装置において、
入力された音声信号を音声特徴情報に変換する音声変換手段と、
該音声変換手段から出力される音声特徴情報に基づいて予め記憶しているカテゴリ群から入力音声のカテゴリを決定するカテゴリ化手段と、
該カテゴリ化手段によりカテゴリ化した入力音声のカテゴリに応じて予め記憶している応答カテゴリを決定する応答カテゴリ決定手段と、
該応答カテゴリ決定手段により決定した応答カテゴリに応じて出力応答文の音声特徴情報を決定する特徴決定手段と、
該特徴決定手段により決定した音声特徴情報に基づいて出力応答文の音声を合成する音声合成手段とを備えることを特徴とする音声対話装置。

【請求項７】　カテゴリ化手段によりカテゴリ化した回数をカテゴリ毎に計数するカテゴリ化計数手段を更に備え、
前記特徴決定手段は、応答カテゴリ決定手段により決定した前記応答カテゴリ及びカテゴリ化計数手段において計数した回数に応じて出力応答文の音声特徴情報を決定する構成としてあることを特徴とする請求項６に記載の音声対話装置。

【請求項８】　カテゴリ化手段によりカテゴリ化した時刻をカテゴリ毎に記憶する時刻記憶手段を更に備え、
前記特徴決定手段は、応答カテゴリ決定手段により決定

した前記応答カテゴリ、カテゴリ化計数手段において計数した回数及び時刻記憶手段に記憶している時刻に応じて出力応答文の音声特徴情報を決定する構成としてあることを特徴とする請求項７に記載の音声対話装置。

【請求項９】　入力された音声に対して応答文を作成して音声出力する音声対話装置において、

入力された音声信号を音声特徴情報に変換する音声変換手段と、

該音声変換手段により出力される音声特徴情報に基づいて予め記憶しているキーワード及び音素群から入力音声のキーワード及び音素を決定するキーワード決定手段と、

該キーワード決定手段において同一キーワードが入力された回数を計数するキーワード計数手段と、

入力されたキーワード及び音素に基づいて応答文を作成する応答文作成手段と、

該応答文作成手段により作成した応答文のうち前記キーワードに対応して作成した応答単語について、キーワードの入力回数に応じて出力される応答文中の各応答単語の韻律又はスペクトルを決定する韻律スペクトル決定手段と、

該韻律スペクトル決定手段により決定した韻律又はスペクトルに基づいて出力応答文の音声を合成する音声合成手段とを備えることを特徴とする音声対話装置。

【請求項１０】　入力された音声に対して応答文を作成して音声出力する音声対話装置において、

入力される音声信号を音声特徴情報に変換する音声変換手段と、

該音声変換手段により出力される音声特徴情報に基づいて予め記憶しているキーワード及び音素群から入力音声のキーワード及び音素を決定するキーワード決定手段と、

該キーワード決定手段において同一キーワードが入力された回数を計数するキーワード計数手段と、

入力されたキーワード及び音素に基づいて応答文を作成する応答文作成手段と、該応答文作成手段により作成した応答文のうち前記キーワードに対応して作成した応答単語について、キーワード入力回数に応じて出力される応答文中の各応答単語の韻律又はスペクトルを決定する韻律スペクトル決定手段と、

前記音声変換手段から出力される音声特徴情報に基づいて予め記憶しているカテゴリ群から入力音声のカテゴリを決定するカテゴリ化手段と、

前記カテゴリ化手段によりカテゴリ化した回数をカテゴリ毎に計数するカテゴリ化計数手段と、

前記カテゴリ化手段によりカテゴリ化した時刻をカテゴリ毎に記憶する時刻記憶手段と、

前記カテゴリ化手段によりカテゴリ化した入力音声のカテゴリに応じて予め記憶している応答カテゴリを決定する応答カテゴリ決定手段と、

該応答カテゴリ決定手段により決定した前記応答カテゴリ、カテゴリ化計数手段により計数した回数及び時刻記憶手段に記憶している時刻に応じて出力応答文の音声特徴情報を決定する特徴決定手段と、

該特徴決定手段により決定した音声特徴情報及び前記韻律スペクトル決定手段により決定した韻律又はスペクトルに基づいて出力応答文の音声を合成する音声合成手段とを備えることを特徴とする音声対話装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、人間とコンピュータとの間の情報伝達を円滑に行うための音声対話方法及び音声対話装置に関する。

【０００２】

【従来の技術】音声を用いてコンピュータと対話を行いながら制御を行う音声対話装置は、カーナビゲーション装置又はゲーム機に採用されておりその用途は拡大している。

【０００３】従来の音声対話装置として特開平７－２１０１９３号が公知であるが、出力される合成音声は単調で、声質、速度、口調又は抑揚等を変えるものではないため、ユーザが機械操作に飽きるという問題があった。また、特開平６－１１０６５０号公報には、入力音声の一部を記憶しておき、音声出力の際はその記憶している音声の速度を変える音声対話装置が開示されている。しかしながらこれらいずれの音声対話装置も話者の感情又は対話の進行状況等を考慮して応答音声を出力するものではなくユーザーフレンドリなものとはいえなかった。例えばカーナビゲーション装置に内蔵されている音声対話装置を使用している場合、ユーザが目的地をうまく見つけることができず次第にいらいらしてくる、又は渋滞により予定時刻に遅れて落胆する等、話者の感情は状況に応じて変化するにも拘わらず従来の音声対話装置はいつも同じ口調で応答するだけであった。

【０００４】また、従来の出力音声は、メリハリが無く一定の口調でなされるため、重要なポイントの把握が困難であるという問題があった。たとえば、走行中にレストランを検索している場合、まず音声対話装置は「３Ｋｍ先左側にＡすし店があります。」と応答したとする。そして、ユーザは「他には？」と音声入力すると、従来のカーナビゲーション装置は同じ声質、速度、口調及び抑揚等で「５Ｋｍ先右側にＢファミリーレストランがあります。」と応答する。この場合、重要な情報は入力音声「他に」に対応する「Ｂファミリーレストラン」であるが、他の情報（「５Ｋｍ」、「先」、「右側」及び「あります」）についても同じ口調で応答するため、重要な情報を把握しづらいという問題があった。特に、音声対話装置と人間との間で交わされる会話内容が長くなると、出力される情報も複雑となり、かかる弊害は顕著になるといえる。

【０００５】
【発明が解決しようとする課題】本発明は斯かる事情に鑑みてなされたものでありその目的とするところは、会話中の重要なポイントを明確化及び話者の感情を認識して応答するユーザーフレンドリな音声対話方法及び音声対話装置を提供することにある。
【０００６】また本発明の他の目的は、対話の進行状況に応じてめまぐるしく変化する話者の感情に柔軟に対応して応答することができるユーザーフレンドリな音声対話方法及び音声対話装置を提供することにある。
【０００７】さらに本発明の他の目的は、出力される応答文中、重要なポイントをユーザが認識しやすいユーザーフレンドリな音声対話方法及び音声対話装置を提供することにある。
【０００８】
【課題を解決するための手段】第１発明に係る音声対話方法は、入力された音声に対して応答文を作成して音声出力する音声対話方法において、入力音声を音声信号に変換するステップと、該音声信号を音声特徴情報に変換するステップと、該音声特徴情報に基づいて所定のカテゴリ群から入力音声のカテゴリを決定するステップと、前記入力音声のカテゴリに応じて所定の応答カテゴリを決定するステップと、該応答カテゴリに応じて出力応答文の音声特徴情報を決定する特徴情報決定ステップと、該特徴情報決定ステップにより決定した音声特徴情報に基づいて出力応答文の音声を合成するステップとを備えることを特徴とする。
【０００９】第２発明に係る音声対話方法は、請求項１に記載の音声対話方法において、カテゴリを決定した回数をカテゴリ毎に計数するステップを更に備え、前記特徴情報決定ステップは、前記応答カテゴリ及び前記ステップで計数した回数に応じて決定することを特徴とする。
【００１０】第３発明に係る音声対話方法は、請求項２に記載の音声対話方法において、カテゴリを決定した時刻をカテゴリ毎に記憶するステップを更に備え、前記特徴決定ステップは、前記応答カテゴリ、前記ステップで計数した回数及び前記ステップで記憶した時刻に応じて決定することを特徴とする。
【００１１】第４発明に係る音声対話方法は、入力された音声に対して応答文を作成して音声出力する音声対話方法において、入力音声を音声信号に変換するステップと、該音声信号を音声特徴情報に変換するステップと、該音声特徴情報に基づいて所定のキーワード及び音素群から入力音声のキーワード及び音素を決定するステップと、同一キーワードが入力された回数を計数するステップと、入力されたキーワード及び音素に基づいて応答文を作成するステップと、作成した応答文のうち前記キーワードに対応して作成した応答単語について、キーワード入力回数に応じて出力する応答文中の各応答単語の韻律又はスペクトルを決定するステップと、前記決定した出力応答文中の応答単語の韻律又はスペクトルに基づいて出力応答文の音声を合成するステップとを備えることを特徴とする。
【００１２】第５発明に係る音声対話方法は、入力された音声に対して応答文を作成して音声出力する音声対話方法において、入力信号を音声信号に変換するステップと、該音声信号を音声特徴情報に変換するステップと、該音声特徴情報に基づいて所定のキーワード及び音素群から入力音声のキーワード及び音素を決定するステップと、同一キーワードが入力された回数を計数するステップと、入力されたキーワード及び音素に基づいて応答文を作成するステップと、作成した応答文のうち前記キーワードに対応して作成した応答単語について、キーワード入力回数に応じて出力する応答文中の各応答単語の韻律又はスペクトルを決定するステップと、前記音声特徴情報に基づいて所定のカテゴリ群から入力音声のカテゴリを決定するステップと、カテゴリを決定した回数をカテゴリ毎に計数するステップと、カテゴリを決定した時刻をカテゴリ毎に記憶するステップと、前記入力音声のカテゴリに応じて所定の応答カテゴリを決定するステップと、前記応答カテゴリ、カテゴリを決定した前記ステップで計数した回数及び前記ステップで記憶した時刻に応じて出力応答文の音声特徴情報を決定するステップと、該決定した出力応答文の音声特徴情報及び前記決定した出力応答文中の応答単語の韻律又はスペクトルに基づいて出力応答文の音声を合成するステップとを備えることを特徴とする。
【００１３】第６発明に係る音声対話装置は、入力された音声に対して応答文を作成して音声出力する音声対話装置において、入力された音声信号を音声特徴情報に変換する音声変換手段と、該音声変換手段から出力される音声特徴情報に基づいて予め記憶しているカテゴリ群から入力音声のカテゴリを決定するカテゴリ化手段と、該カテゴリ化手段によりカテゴリ化した入力音声のカテゴリに応じて予め記憶している応答カテゴリを決定する応答カテゴリ決定手段と、該応答カテゴリ決定手段により決定した応答カテゴリに応じて出力応答文の音声特徴情報を決定する特徴決定手段と、該特徴決定手段により決定した音声特徴情報に基づいて出力応答文の音声を合成する音声合成手段とを備えることを特徴とする。
【００１４】第７発明に係る音声対話装置は、請求項６に記載の音声対話装置において、カテゴリ化手段によりカテゴリ化した回数をカテゴリ毎に計数するカテゴリ化計数手段を更に備え、前記特徴決定手段は、応答カテゴリ決定手段により決定した前記応答カテゴリ及びカテゴリ化計数手段において計数した回数に応じて出力応答文の音声特徴情報を決定する構成としてあることを特徴とする。
【００１５】第８発明に係る音声対話装置は、請求項７

に記載の音声対話装置において、カテゴリ化手段によりカテゴリ化した時刻をカテゴリ毎に記憶する時刻記憶手段を更に備え、前記特徴決定手段は、応答カテゴリ決定手段により決定した前記応答カテゴリ、カテゴリ化計数手段において計数した回数及び時刻記憶手段に記憶している時刻に応じて出力応答文の音声特徴情報を決定する構成としてあることを特徴とする。

【００１６】第９発明に係る音声対話装置は、入力された音声に対して応答文を作成して音声出力する音声対話装置において、入力された音声信号を音声特徴情報に変換する音声変換手段と、該音声変換手段により出力される音声特徴情報に基づいて予め記憶しているキーワード及び音素群から入力音声のキーワード及び音素を決定するキーワード決定手段と、該キーワード決定手段において同一キーワードが入力された回数を計数するキーワード計数手段と、入力されたキーワード及び音素に基づいて応答文を作成する応答文作成手段と、該応答文作成手段により作成した応答文のうち前記キーワードに対応して作成した応答単語について、キーワードの入力回数に応じて出力される応答文中の各応答単語の韻律又はスペクトルを決定する韻律スペクトル決定手段と、該韻律スペクトル決定手段により決定した韻律又はスペクトルに基づいて出力応答文の音声を合成する音声合成手段とを備えることを特徴とする。

【００１７】第１０発明に係る音声対話装置は、入力された音声に対して応答文を作成して音声出力する音声対話装置において、入力される音声信号を音声特徴情報に変換する音声変換手段と、該音声変換手段により出力される音声特徴情報に基づいて予め記憶しているキーワード及び音素群から入力音声のキーワード及び音素を決定するキーワード決定手段と、該キーワード決定手段において同一キーワードが入力された回数を計数するキーワード計数手段と、入力されたキーワード及び音素に基づいて応答文を作成する応答文作成手段と、該応答文作成手段により作成した応答文のうち前記キーワードに対応して作成した応答単語について、キーワード入力回数に応じて出力される応答文中の各応答単語の韻律又はスペクトルを決定する韻律スペクトル決定手段と、前記音声変換手段から出力される音声特徴情報に基づいて予め記憶しているカテゴリ群から入力音声のカテゴリを決定するカテゴリ化手段と、前記カテゴリ化手段によりカテゴリ化した回数をカテゴリ毎に計数するカテゴリ化計数手段と、前記カテゴリ化手段によりカテゴリ化した時刻をカテゴリ毎に記憶する時刻記憶手段と、前記カテゴリ化手段によりカテゴリ化した入力音声のカテゴリに応じて予め記憶している応答カテゴリを決定する応答カテゴリ決定手段と、該応答カテゴリ決定手段により決定した前記応答カテゴリ、カテゴリ化計数手段により計数した回数及び時刻記憶手段に記憶している時刻に応じて出力応答文の音声特徴情報を決定する特徴決定手段と該特徴決定手段により決定した音声特徴情報及び前記韻律スペクトル決定手段により決定した韻律又はスペクトルに基づいて出力応答文の音声を合成する音声合成手段とを備えることを特徴とする。

【００１８】第１発明及び第６発明にあっては、入力音声の速度、抑揚、パワー及びスペクトル等の音声特徴情報を抽出する。そして、「疑問」、「落胆」又は「いらだち」等のカテゴリ毎に記憶している音声特徴情報群と入力音声の音声特徴情報とを比較する。比較した結果、最も近似する音声特徴情報に対応するカテゴリを入力音声のカテゴリと決定する。さらに、「親切な応答」、「速い応答」又は「丁寧な応答」等の応答カテゴリ群を予め記憶しておき入力音声のカテゴリに応じて適切な一の応答カテゴリを決定する。そして、応答カテゴリの種類に応じて出力応答文の音声速度、抑揚、パワー及びスペクトル等の音声特徴情報を合成するようにしたので、例えば、話者が急いで問いかければ速い速度で応答するといったことが可能となり単調さが無く、あたかも人間と会話している様なユーザーフレンドリな音声対話システムの構築が可能となる。

【００１９】第２発明及び第７発明にあっては、カテゴリ化した回数を計数し、その回数に応じて出力応答文の音声速度、抑揚、パワー及びスペクトル等の音声特徴情報を合成するようにしたので、より単調さが無くなり、対話の中で刻々と変化する話者の感情をも考慮したユーザーフレンドリな音声対話システムの構築が可能となる。

【００２０】第３発明及び第８発明にあっては、カテゴリ化した回数及び時刻を記憶しておきそのカテゴリ化回数及び時刻に応じて、出力応答文の音声速度、抑揚、パワー及びスペクトル等の音声特徴情報を合成するようにしたので、さらに単調さが無くなり、話者の心境変化を経時的に考慮したよりユーザーフレンドリな音声対話システムの構築が可能となる。

【００２１】第４発明及び第９発明にあっては、同じキーワードが入力される回数を計数する。そして入力頻度の高いキーワードに対応する応答単語ついては、韻律又はスペクトルを変更する。例えば入力頻度の高いキーワードに対応する応答単語については、入力頻度の低いキーワードに対応する応答単語よりも、音量を大きくして出力させるようにしたので、話者は応答音声中重要なポイントを的確に聞くことが可能となる。

【００２２】第５発明及び第１０発明にあっては、同じキーワードが入力される回数を計数する。また、入力音声カテゴリに対する応答カテゴリを決定すると共に、カテゴリ化した回数を計数、及びカテゴリ化した時刻を記憶する。そして、応答カテゴリ、カテゴリ化した回数、カテゴリ化した時刻及びキーワード入力回数に基づいて、出力応答文の音声特徴情報及び出力応答文中の各応答単語の韻律又はスペクトルを合成するようにしたの

で、単調さが無く、また出力音声の重要なポイントを聞き取りやすいユーザーフレンドリな音声対話システムを構築することができる。

【００２３】

【発明の実施の形態】以下本発明をその実施の形態を示す図面に基づいて詳述する。図１は本発明の音声対話装置Ｄを示すブロック図である。図において１はマイクロフォンであり、入力音声を電気的な音声信号に変換する。この音声信号はＡ／Ｄ変換器４によりデジタル化され、デジタル化された音声信号はＭＰＵ２で音声信号から音声特徴情報への変換、キーワード又は音素の検出、応答文の作成及びカテゴリの決定等の各処理が行われる。また、ＭＰＵ２内部は時計部７を備える。なおＭＰＵ２の処理については後述する。

【００２４】また、ＭＰＵ２にはハードディスク等の記憶装置３が接続されておりキーワードデータ、音素データ、カテゴリーデータ等の各種情報が記憶されている。音声対話装置Ｄと図示しないカーナビゲーション装置又はゲーム装置等とはＩ／Ｏ部８において制御信号、距離情報データ又は渋滞情報テキストファイル等の各種情報を送受信する。ＭＰＵ２で生成された応答文データ、音量データ、音声特徴情報等の各データはＤ／Ａ変換器５により電気信号に変換される。そうすると、電気信号はスピーカ６から音声として出力されることになる。

【００２５】図２は、ＭＰＵ２の処理手順を示すフローチャートである。まず、Ａ／Ｄ変換器４から出力されたデジタルの音声信号を速度、抑揚、パワーまたはＦＦＴによる周波数スペクトル等の音声特徴情報に変換する（ステップＳ２１）。ステップＳ２１によって処理された音声特徴情報は、キーワード等を検出して応答文等を作成するキーワード処理と話者の感情カテゴリを検出するカテゴリ処理とを実行する。

【００２６】まず、カテゴリ処理について以下に説明する。記憶装置３におけるカテゴリデータベース３ｄにはカテゴリ毎（疑問、確認又はいらだち等）に音声特徴情報（例えば入力音声の速度、抑揚又は周波数スペクトル等）が記憶されている。まず、ステップＳ２１において変換された音声特徴情報に基づき入力音声のカテゴリを決定する（ステップＳ２２）。入力音声のカテゴリ決定にあっては入力音声の音声特徴情報とカテゴリデータベース３ｄに記憶している音声特徴情報とを比較し、一致又は最も近似する音声特徴情報に対応するカテゴリを入力音声（話者）のカテゴリとして決定する（ステップＳ２２）。

【００２７】そして、ステップＳ２２においてカテゴリ化したカテゴリ毎にカウンタを設けておき同一カテゴリが入力されるとカウンタをインクリメントする（ステップＳ２３）。また、ステップＳ２３においてカテゴリ化した時刻をカテゴリ毎に逐次記憶する（ステップＳ２４）。時刻の記憶は、時計部７からの信号に基づいて行われ、ステップＳ２２においてカテゴリ化するたびに逐次記憶する。記憶装置３における応答カテゴリデータベース３ｅには、入力音声のカテゴリに対応する応答カテゴリ群が予め記憶されている。この応答カテゴリデータベース３ｅを参酌して、ステップＳ２２においてカテゴリ化した入力音声のカテゴリに対応する応答カテゴリを決定する（ステップＳ２５）。記憶装置３における音調データベース３ｆには応答カテゴリに対応する音声特徴情報が記憶されている。この音調データベース３ｆを参照して、ステップ２５において決定した応答カテゴリに対する音声特徴情報を決定する（ステップＳ２６）。

【００２８】次に、キーワード処理について説明する。記憶装置３のキーワードデータベース３ａにはキーワード毎に音声特徴情報が記憶されており、同様に記憶装置３の音素データベース３ｂには音素毎に音声特徴情報が記憶されている。まず、ステップＳ２１によって処理された音声特徴情報とキーワードデータベース３ａ及び音素データベース３ｂに記憶している音声特徴情報とを対比し、一致する音声特徴情報に対応するキーワード及び音素を決定する（ステップＳ２７）。

【００２９】続いて、ステップ２７において、キーワード毎にカウンタを設けておき、同一キーワードが入力されるとカウンタをインクリメントする（ステップＳ２８）。そして、ステップＳ２７により決定したキーワード及び音素に基づいて応答文を作成する（ステップＳ２９）。記憶装置３の辞書データベース３ｃにはキーワード及び音素の発音記号及び意味内容が記憶されており、応答文作成処理（ステップＳ２９）では辞書データベース３ｃを参照しながらする応答文を作成する。そして、ステップＳ２８においてキーワードを計数した回数に応じて、ステップＳ２９において作成した応答文中のキーワードに対応する応答単語毎に韻律又はスペクトルを決定する（ステップＳ２１０）。なお、韻律とは音の高さ、速さ又は強さ等をいう。

【００３０】上述のキーワード処理及びカテゴリ処理手順が終了した場合はこれら２つの処理結果を応答文の音声に反映させる。つまり、ステップＳ２９において作成した応答文の音声をステップＳ２６において決定した音声特徴情報及びステップＳ２１０で決定したキーワードに対応する各応答単語の韻律又はスペクトル情報に基づいて合成する（ステップＳ２１１）のである。以下に各処理内容について詳述する。なお、キーワード処理及びカテゴリ処理についてはどちらを先に処理しても良い。

【００３１】図３はカテゴリ化処理の手順を表す説明図である。図において、カテゴリデータベース３ｄはカテゴリに毎に音声特徴情報が既定値として記憶されている。音声特徴情報は例えば、速度、抑揚、パワー及びスペクトルである。例えば、「疑問」カテゴリであれば、語尾が上がっているというような情報が記憶されている。

【００３２】カテゴリ化処理（ステップＳ２２）では、入力された音声の音声特徴情報とカテゴリデータベース３ｄのカテゴリ毎に既定値として記憶されている音声特徴情報とをパターンマッチング等の手法により比較する。比較した結果、一致又は最も近似する音声特徴情報に対応してカテゴリデータベース３ｄに記憶しているカテゴリを音声入力のカテゴリとして決定する。なお、カテゴリ決定の精度を向上させるために、ユーザがディクテーションを行い、その結果をカテゴリデータベース３ｄに記憶している音声特徴情報を適宜更新するようにしても良い。さらに、この本発明に係る音声対話装置を複数人で利用する場合は、ユーザ毎のディクテーション結果を分別して記憶するようにする。

【００３３】図４はカテゴリ化処理手順を示すフローチャートである。まず、ステップＳ２１で変換された入力された音声の音声特徴情報とカテゴリデータベース３ｄのカテゴリ毎に既定値として記憶されている音声特徴情報とを比較する（ステップＳ４１）。そして、カテゴリデータベース３ｄの音声特徴群から最適な一の音声特徴情報を決定する（ステップＳ４２）。それから、係る音声特徴情報に対応するカテゴリを入力音声のカテゴリとして決定する（ステップＳ４３）。

【００３４】図５は応答カテゴリ決定の処理手順を示す説明図である。応答カテゴリデータベース３ｅには入力カテゴリに対する応答カテゴリがそれぞれ記憶されている。例えば、入力音声カテゴリが「急ぎ」である場合、出力される音声もきびきびと出力した方がよいので、これに対応する応答カテゴリは「速い応答」となるように記憶されている。また、入力音声カテゴリが「落胆」である場合、ユーザを落ち着かせる出力音声が好ましいので、これに対応する応答カテゴリは「励ます応答」が記憶されている。応答カテゴリ決定処理（ステップＳ２５）では、応答カテゴリデータベース３ｅに記憶している情報を基に、カテゴリ化処理（ステップＳ２２）において決定したカテゴリに対応する応答カテゴリを決定する処理を行う。

【００３５】音調データベース３ｆには、応答カテゴリ毎に応答音声の音声特徴情報が記憶されている。例えば、「速い応答」であれば音声特徴情報の一つである既定速度「Ｃ（やや速いスピード）」が選択され、さらに、既定抑揚「ハ（一定の抑揚）」及び既定パワー「ｃ」が選択される。この音声特徴情報はユーザの設定により既定値を変更することもできる。特徴決定処理（ステップＳ２６）では、音調データベース３ｆを参照して、応答カテゴリ決定処理（ステップＳ２５）により決定した応答カテゴリに基づいて出力される音声の音声特徴情報を決定する処理を行う。

【００３６】図６はカテゴリ化計数処理及び特徴決定処理の手順を示す説明図である。カテゴリ化計数処理（ステップＳ２３）においてはカテゴリ化処理（ステップＳ２２）でカテゴリ化した回数をカテゴリ毎に記憶する処理を行う。図の例であれば、「疑問」カテゴリが「１」回、「いらだち」カテゴリが「３」回及び「急ぎ」カテゴリが「２」回とそれぞれ記憶されている。応答カテゴリ決定処理（ステップＳ２５）においては、入力カテゴリに対する応答カテゴリが決定されるが、同様に入力カテゴリ化回数に対する応答カテゴリの回数も計数する。図の例であれば、入力カテゴリ「疑問」に対する応答カテゴリ「親切な応答」が「１」回、入力カテゴリ「いらだち」に対する応答カテゴリ「丁寧な応答」が「３」回及び入力カテゴリ「急ぎ」に対する応答カテゴリ「速い応答」が「２」回とそれぞれ計数されている。

【００３７】特徴決定処理（ステップＳ２６）においては、ステップＳ２５において決定した応答カテゴリ及び回数を考慮して音声特徴情報を決定する処理を行う。例えば、「親切な応答」を連続して行う場合は、１回目よりも２回目の方が「より親切な応答」になるように制御し、あるいは、「丁寧な応答」が続いているような場合は、「より丁寧な応答」になるように応答音声の音声特徴を制御する。あるいは、同一応答カテゴリが連続していない場合でも、カウントされた回数に応じて、柔軟に制御を行う。

【００３８】図７は時刻記憶処理の手順を示す説明図である。時刻記憶処理（ステップＳ２４）においては、カテゴリ化処理（ステップＳ２２）においてカテゴリ化した時刻をカテゴリ毎に逐次記憶する処理を行う。そして、記憶した時刻を考慮して音声特徴情報を決定する。例えば、カテゴリが「いらだち」と決定した後、次に入力された音声のカテゴリが「落胆」であるとする。この場合、時系列的に入力音声は「いらだち」から「落胆」というカテゴリに変化したことを考慮し、応答カテゴリとして「丁寧かつ励ましの応答」等といったカテゴリを選択する。これを受けて特徴決定処理（ステップＳ２６）においては、上述のカテゴリ化計数処理（ステップＳ２３）と同じ手法により音声特徴情報を決定する。そうすると、経時的要素をも考慮した応答音声が出力されることになる。

【００３９】カテゴリ化計数処理（ステップＳ２３）における計数回数及び時刻記憶処理（ステップＳ２４）により記憶している時刻は、図示しない音声対話装置の制御部からリセット要求があるまで随時更新される。例えばカーナビゲーション装置においては、一つのタスクが終了するまで（例えば、目的地を検索している場合はその検索が終了するまで）、音声対話が連続して行われることになるため、その間は回数及び時刻をクリアしないでおく。逆に、他のタスクへ移行（例えば、目的地の検索が終了して渋滞情報の案内へ移行）する場合は、対話内容が異なるので回数及び時刻をクリアして初期化する。このように回数及び時刻を「０」に更新する要求がある場合は回数及び時刻を「０」にクリアする。

【００４０】図８はカテゴリ化回数及び時刻の更新処理手順を示すフローチャートである。まず、カテゴリを決定、カテゴリ化した回数を計数、及びカテゴリ化した時刻を記憶する（ステップＳ８１）。そして、入力音声カテゴリに基づいて応答カテゴリを決定する（ステップＳ８２）。ついで、応答カテゴリ、カテゴリ化回数及びカテゴリ化時刻に基づいて音声特徴情報を決定する（ステップＳ８３）。そして、図示しない制御部から回数及び時刻のクリア要求がない場合は（ステップＳ８４でＮｏ）、ステップＳ８１へ戻る。

【００４１】そして、再度音声が入力されカテゴリ化した場合は、カテゴリに応じて計数している回数をインクリメントする（ステップＳ８１）。同時に、カテゴリ化した時刻もカテゴリ毎に逐次追加記憶する（ステップＳ８１）。このとき、計数値及び記憶時刻に応じて重み付けをして、音声特徴情報を決定する（ステップＳ８３）。一方、図示しない制御部から回数及び時刻のクリア要求がある場合は（ステップＳ８４でＹｅｓ）、カテゴリ化回数を「０」にすると共に、記憶している時刻を全て消去する（ステップＳ８５）。

【００４２】図９は本発明のキーワード処理の手順を示す説明図である。キーワード決定処理（ステップＳ２７）では、キーワードを検出するたびにキーワード毎に計数する処理を行う（ステップＳ２８）。例えば、カーナビゲーション装置において、ユーザがコンビニエンスストアを検索する場合に、ユーザが「コンビニ」と発声すると、キーワード「コンビニ」を「１」と計数する。なお、キーワード計数回数についてもカテゴリ回数と同様に、図示しない制御部からクリア要求があるまで、回数を蓄積する。

【００４３】応答文作成処理（ステップＳ２９）では、キーワードに基づいて、辞書データベース３ｃ及び図示しないカーナビゲーション装置の地図データベースへアクセスしてコンビニを検索し応答文を作成する処理を行う。そして「３Ｋｍ先右側にハイソンがあります。」という応答文を作成する。韻律スペクトル決定処理（ステップＳ２１０）では、応答文中、キーワード「コンビニ」に対応する応答単語「ハイソン」をキーワードの計数回数に基づいて「１」と重み付けする。一方、他の応答単語「３Ｋｍ先右側に」及び「があります。」についてはキーワードの計数回数は「０」なので、重みを「０」とする処理を行う。そして、この重みに基づいて韻律又はスペクトルを決定する処理を行う。例えば、応答単語「３Ｋｍ先右側に」及び「があります。」は音量を通常の音量とし、応答単語「ハイソン」は通常の音量よりも高い音量となるよう決定する。

【００４４】更に対話が継続するのであれば、入力されるキーワードを更に計数する。具体的に示すと、続いて「他は？」と入力されるとキーワード「他は？」が「１」と計数される。そして、応答文が「他は近くにありません。」であるとすると、キーワード「コンビニ」及び「他は？」について回数は共に「１」と計数されているので、キーワード「他は？」に対応する応答単語「他は」及びキーワード「コンビニ」に対応する応答単語「ありません。」を同じ重みとし、応答単語「他は」及び「ありません」を同じ音量で出力する。一方、他の応答単語「近くに」は計数回数が「０」つまり重みが「０」であるので、応答単語「他は」及び「ありません」よりも低い音量となるよう設定する。

【００４５】さらにユーザが「他は？どこ？」と発声したとする。すると、キーワードは「コンビニ」が「１」回、「他は？」が「２」回、「どこ？」が「１」回とそれぞれ計数される。この場合応答文が「ハイソンの他は、１０Ｋｍ先左側にセブンがあります。それ以外は１０Ｋｍ以上離れていますねぇ。」であるとする。そうすると、検出回数が多いものから順に、キーワード「他は？」の応答単語「の他は」の重みが「２」に、キーワード「コンビニ」の応答単語「ハイソン」及び「セブン」の重みが「１」に、これと同じ計数回数であるキーワード「どこ？」の応答単語「１０Ｋｍ先左側」及び「１０Ｋｍ以上離れて」が重み「１」に、そしてそれ以外の応答単語「に」、「があります。それ以外は」及び「いますねぇ。」の重みは「０」とそれぞれ計数回数に応じて重み付けされる。これにより、かかる重みに基づいて、重量順に音量を決定する。

【００４６】図１０は、音声合成処理手順を示す説明図である。上記例で、従来の音声対話装置では、何の強調、抑揚、感情等もなく一律に作成された応答文「ハイソンの他は、１０Ｋｍ先左側にセブンがあります。それ以外は１０Ｋｍ以上離れていますねぇ。」を出力するものであった。これに対し本発明は、韻律スペクトル決定処理（ステップＳ２１０）により前記重みに基づいて応答単語毎に韻律又はスペクトルを決定する処理を行う。一方、応答カテゴリ決定処理（ステップＳ２５）により決定した応答カテゴリ（図の例では「親切な応答」）に対応する音声特徴情報を特徴決定処理（ステップＳ２６）により決定する（図の例では速度「ゆっくり」及び抑揚「語尾下げる」と決定する）。そうすると、韻律スペクトル決定処理（ステップＳ２１０）で決定した韻律又はスペクトル及び特徴決定処理（ステップＳ２６）で決定した音声特徴情報を基に、応答文作成処理（ステップＳ２９）で作成した応答文の音声データに、音声特徴情報及び韻律又はスペクトル情報が付加する処理が行われ（音声合成処理ステップＳ２１１）、重要なポイントを聞き取りやすく、しかも感情を持ったような応答音声が出力されることになる。

【００４７】

【発明の効果】以上詳述した如く第１発明及び第６発明にあっては、入力音声の速度、抑揚、パワー及びスペクトル等の音声特徴情報を抽出する。そして、「疑問」、

15

「落胆」又は「いらだち」等のカテゴリ毎に記憶している音声特徴情報群と入力音声の音声特徴情報とを比較する。比較した結果、最も近似する音声特徴情報に対応するカテゴリを入力音声のカテゴリと決定する。さらに、「親切な応答」、「速い応答」又は「丁寧な応答」等の応答カテゴリ群を予め記憶しておき入力音声のカテゴリに応じて適切な一の応答カテゴリを決定する。そして、応答カテゴリの種類に応じて出力応答文の音声速度、抑揚、パワー及びスペクトル等の音声特徴情報を合成するようにしたので、例えば、話者が急いで問いかければ速い速度で応答するといったことが可能となり単調さが無く、あたかも人間と会話している様なユーザーフレンドリな音声対話システムの構築が可能となる。

【００４８】また、第２発明及び第７発明にあっては、カテゴリ化した回数を計数し、その回数に応じて出力応答文の音声速度、抑揚、パワー及びスペクトル等の音声特徴情報を合成するようにしたので、より単調さが無くなり、対話の中で刻々と変化する話者の感情をも考慮したユーザーフレンドリな音声対話システムの構築が可能となる。

【００４９】また、第３発明及び第８発明にあっては、カテゴリ化した回数及び時刻を記憶しておきそのカテゴリ化回数及び時刻に応じて、出力応答文の音声速度、抑揚、パワー及びスペクトル等の音声特徴情報を合成するようにしたので、さらに単調さが無くなり、話者の心境変化を経時的に考慮したよりユーザーフレンドリな音声対話システムの構築が可能となる。

【００５０】また、第４発明及び第９発明にあっては、同じキーワードが入力される回数を計数する。そして入力頻度の高いキーワードに対応する応答単語ついては韻律又はスペクトルを変化、例えば入力頻度の低いキーワードに対応する応答単語よりも、音量を大きくして出力させるようにしたので、話者は応答音声中重要なポイントを的確に聞くことが可能となる。

【００５１】さらに、第５発明及び第１０発明にあっては、同じキーワードが入力される回数を計数する。また、入力音声カテゴリに対する応答カテゴリを決定すると共に、カテゴリ化した回数を計数、及びカテゴリ化した時刻を記憶する。そして、応答カテゴリ、カテゴリ化した回数、カテゴリ化した時刻及びキーワード入力回数に基づいて、出力応答文の音声特徴情報及び出力応答文中の各応答単語の韻律又はスペクトルを合成するようにしたので、単調さが無く、また出力音声の重要なポイントを聞き取りやすいユーザーフレンドリな音声対話システムを構築することができる。

16

【図面の簡単な説明】

【図１】本発明の音声対話装置を示すブロック図である。

【図２】ＭＰＵの処理手順を示すフローチャートである。

【図３】カテゴリ化処理の手順を表す説明図である。

【図４】カテゴリ化処理手順を示すフローチャートである。

【図５】応答カテゴリ決定の処理手順を示す説明図である。

【図６】カテゴリ化計数処理及び音調決定処理の手順を示す説明図である。

【図７】時刻記憶処理の手順を示す説明図である。

【図８】カテゴリ化回数及び時刻の更新処理手順を示すフローチャートである。

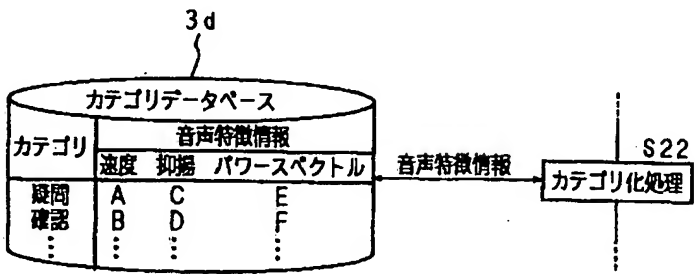
【図９】本発明のキーワード処理の手順を示す説明図である。

【図１０】音声合成処理手順を示す説明図である。

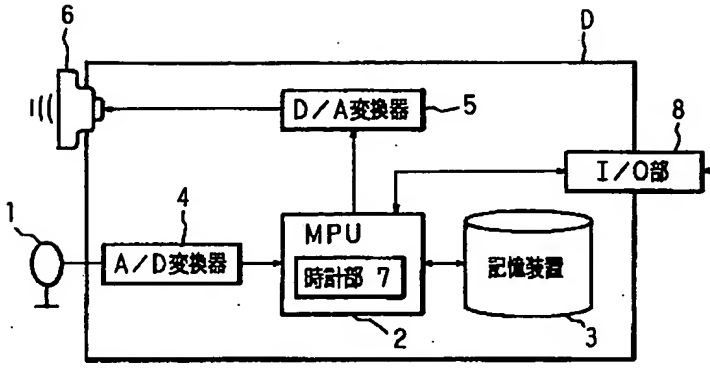
【符号の説明】

Ｄ　音声対話装置
１　マイクロフォン
２　ＭＰＵ
３　記憶装置
３ａ　キーワードデータベース
３ｂ　音素データベース
３ｃ　辞書データベース
３ｄ　カテゴリデータベース
３ｅ　応答カテゴリデータベース
３ｆ　音調データベース
６　スピーカ

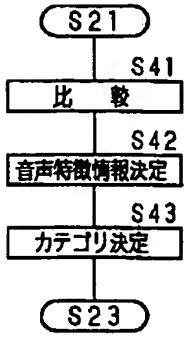
【図３】



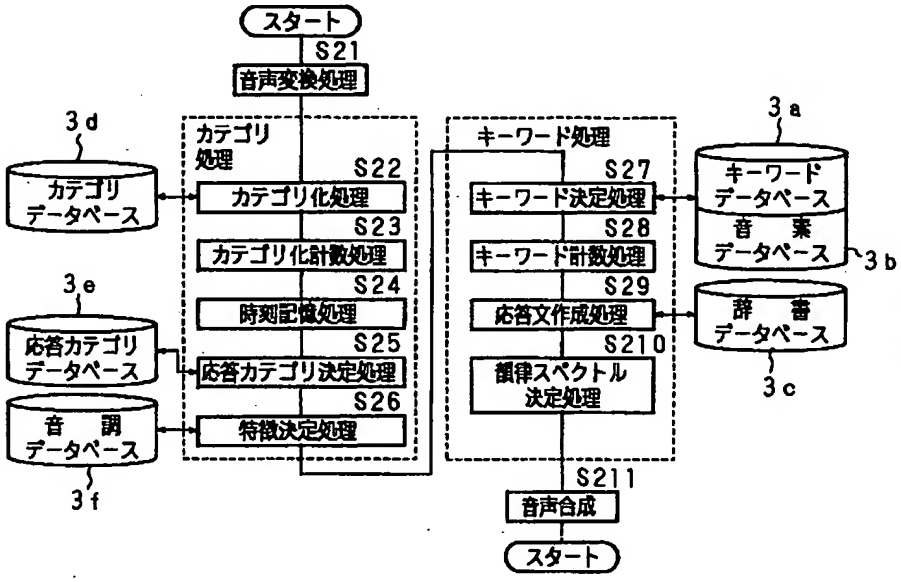
【図１】



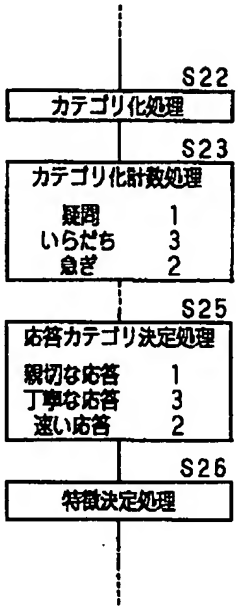
【図４】



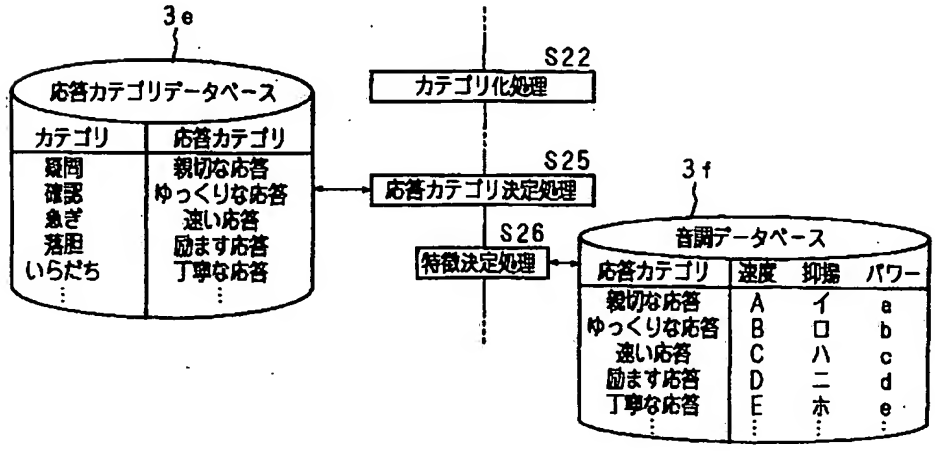
【図２】



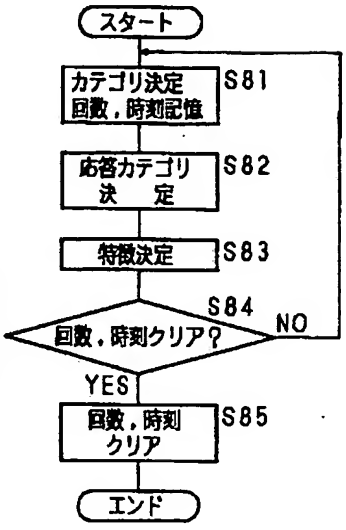
【図６】



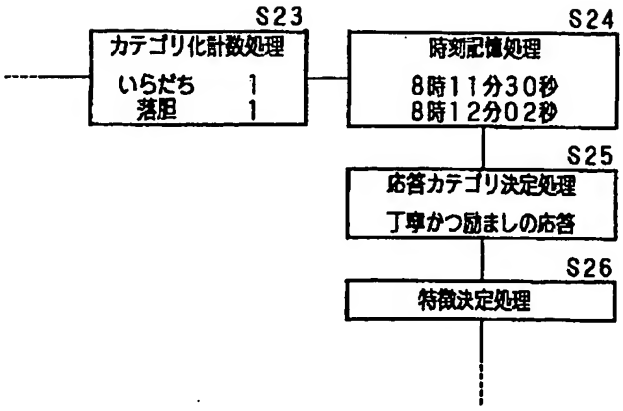
【図５】



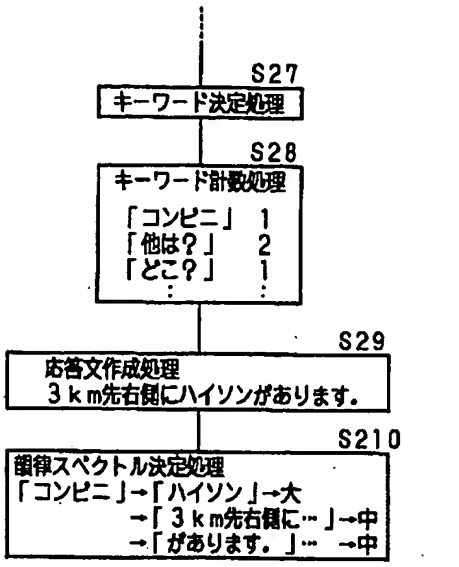
【図８】



【図７】

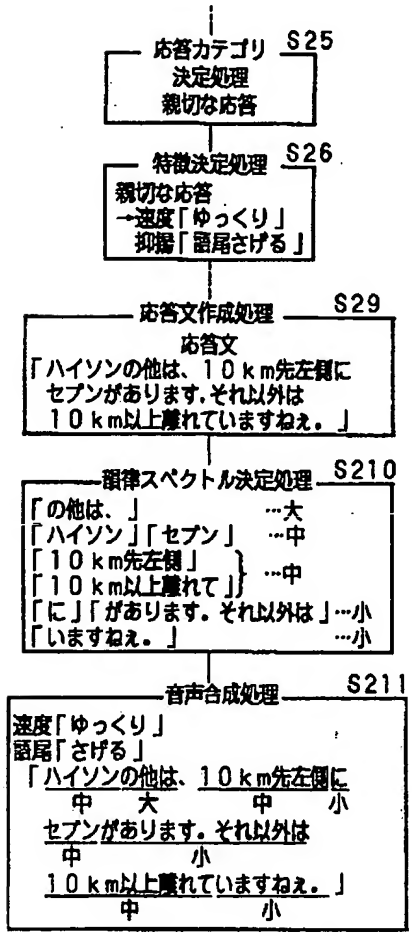


【図９】



会話例　Ｄ：音声対話装置　Ｕ：ユーザ
Ｕ：コンビニ
Ｄ：３ｋｍ先にハイソンがあります。
Ｕ：他は？
Ｄ：他は近くにありません。
Ｕ：他は？どこ？
Ｄ：ハイソンの他は、１０ｋｍ先左側にセブンがあります。それ以外は１０ｋｍ以上離れていますねぇ。

【図１０】



フロントページの続き

(51) Int.Cl.[7]	識別記号	F I	ﾃｰﾏｺｰﾄﾞ (参考)
15/22			

Japanese Laid-open Patent
Laid-open Number: 2001-272991
Laid-open Date: October 5, 2001
Application Number: 2000-84948
Filing Date: March 24, 2000
Applicant: Sanyo Electric Co., Ltd.

[Title of the Invention] VOICE INTERACTIVE METHOD AND VOICE INTERACTIVE DEVICE

[Abstract]

[Problem] An apparatus equipped with a voice interactive function such as a car navigation system or a TV game apparatus appears. However, no apparatus appears capable of changing the voice quality and tone of an output voice in correspondence to a progress of voice interaction, and a synthetic voice is outputted in monotonous voice quality and tone. Therefore, a problem arises in that a user loses interest in an operation of a machine.

[Solving Means] A feeling of a speaker is categorized by analyzing a speed, intonation, or the like of an input voice and a response voice is also categorized in correspondence to that category to make a voice interactive system respond in a feeling-like manner. Hence, it is possible to construct a user-friendly voice interactive system which is free from the monotony and with which a user talks as if he/she has a conversation with a human. Moreover, a response voice for a response word corresponding to a keyword having a high input frequency is outputted with more emphasis than that

corresponding to a keyword having a low input frequency. Hence, a speaker can exactly catch an important point in the response voice.

[Scope of Claims for a Patent]

[Claim 1] A voice interactive method of making a response statement in response to an inputted voice to output a voice, comprising:

a step of converting an input voice into a voice signal;

a step of converting the voice signal into voice feature information;

a step of determining a category of the input voice from a predetermined category group based on the voice feature information;

a step of determining a predetermined response category in correspondence to the category of the input voice;

a step of determining voice feature information of an output response statement in correspondence to the response category; and

a step of obtaining a voice of the output response statement through synthesis based on the voice feature information determined in the feature information determining step.

[Claim 2] A voice interactive method according to claim 1, further comprising a step of counting the number of times on a category basis by which the categories are determined,

wherein the feature information in the feature information determining step is determined in correspondence to the response category and the number of times counted in the step of counting the number of times.

[Claim 3] A voice interactive method according to claim 2, further comprising a step of storing information on time on a category basis

when the categories are determined,

wherein the feature information in the feature information determining step is determined in correspondence to the number of times counted in the step of counting the number of times and the information on the time stored in the step of storing information on time.

[Claim 4] A voice interactive method of making a response statement in response to an inputted voice to output a voice, comprising:

a step of converting an input voice into a voice signal;

a step of converting the voice signal into voice feature information;

a step of determining a keyword and a phoneme of the input voice from a predetermined keyword and phoneme group based on the voice feature information;

a step of counting the number of times by which the same keyword is inputted;

a step of making a response statement based on the inputted keyword and phoneme;

a step of, with respect to response words of the made response statement which are made in correspondence to the keyword, determining a rhythm or a spectrum of each of the response words in the response statement to be outputted in correspondence to the number of times of keyword inputs; and

a step of obtaining a voice for the output response statement

through synthesis based on the determined rhythms or spectra of the response words in the output response statement.

[Claim 5] A voice interactive method of making a response statement in response to an inputted voice to output a voice, comprising:

a step of converting an input voice into a voice signal;

a step of converting the voice signal into voice feature information;

a step of determining a keyword and a phoneme of the input voice from a predetermined keyword and phoneme group based on the voice feature information;

a step of counting the number of times by which the same keyword is inputted;

a step of making a response statement based on the inputted keyword and phoneme;

a step of, with respect to response words of the made response statement which are made in correspondence to the keyword, determining a rhythm or a spectrum of each of the response words in the response statement to be outputted in correspondence to the number of times of keyword inputs;

a step of determining a category of the input voice from a predetermined category group based on the voice feature information;

a step of counting the number of times on a category basis by which the categories are determined;

a step of storing information on time on a category basis when

the categories are determined;

a step of determining a predetermined response category in correspondence to the category of the input voice;

a step of determining voice feature information of the output response statement in correspondence to the response category, the number of times by which the categories are determined counted in the step of counting the number of times, and the time the information on which is stored in the time storing step; and

a step of obtaining a voice for the output response statement through synthesis based on the determined voice feature information of the output response statement and the determined rhythms or spectra of the response words in the output response statement.

[Claim 6] A voice interactive device for making a response statement in response to an inputted voice to output a voice, comprising:

voice converting means for converting an inputted voice signal into voice feature information;

categorizing means for determining a category of the input voice from a category group information on which is stored in advance based on the voice feature information outputted from the voice converting means;

response category determining means for determining a response category information on which is stored in advance in correspondence to the category of the input voice categorized by the categorizing means;

feature determining means for determining voice feature information of an output response statement in correspondence to the response category determined by the response category determining means; and

voice synthesizing means for obtaining a voice for the output response statement through synthesis based on the voice feature information determined by the feature determining means.

[Claim 7] A voice interactive device according to claim 6, further comprising categorization counting means for counting the number of times on a category basis by which the categorization is carried out by the categorizing means,

wherein the feature determining means determines voice feature information of an output response statement in correspondence to the response category determined by the response category determining means and the number of times counted by the categorization counting means.

[Claim 8] A voice interactive device according to claim 7, further comprising time storing means for storing information on time on a category basis when the categorization is carried out by the categorizing means,

wherein the feature determining means determines voice feature information of an output response statement in correspondence to the response category determined by the response category determining means, the number of times counted by the categorization

counting means, and the time the information on which is stored in the time storing means.

[Claim 9] A voice interactive device for making a response statement in response to an inputted voice to output a voice, comprising:

voice converting means for converting an inputted voice signal into voice feature information;

keyword determining means for determining a keyword and a phoneme of an input voice from a keyword and phoneme group information on which is stored in advance based on the voice feature information outputted from the voice converting means;

keyword counting means for counting the number of times by which the same keyword is inputted in the keyword determining means;

response statement making means for making a response statement based on the inputted keyword and phoneme;

rhythm/spectrum determining means for, with respect to response words, of the response statement made by the response statement making means, which is made in correspondence to the keyword, determining a rhythm or a spectrum of each of response words in the response statement outputted in correspondence to the number of times of keyword inputs; and

voice synthesizing means for obtaining a voice of the output response statement through synthesis based on the rhythm or spectrum determined by the rhythm/spectrum determining means.

[Claim 10] A voice interactive device for making a response statement

in response to an inputted voice to output a voice, comprising:

voice converting means for converting an inputted voice signal into voice feature information;

keyword determining means for determining a keyword and a phoneme of an input voice from a keyword and phoneme group information on which is stored in advance based on the voice feature information outputted from the voice converting means;

keyword counting means for counting the number of times by which the same keyword is inputted in the keyword determining means;

response statement making means for making a response statement based on the inputted keyword and phoneme;

rhythm/spectrum determining means for, with respect to response words, of the response statement made by the response statement making means, which is made in correspondence to the keyword, determining a rhythm or a spectrum of each of response words in the response statement outputted in correspondence to the number of times of keyword inputs;

categorizing means for determining a category of the input voice from a category group information on which is stored in advance based on the voice feature information outputted from the voice converting means;

categorization counting means for counting the number of times on a category basis by which the categorization is carried out by the categorizing means;

time storing means for storing information on time on a category basis when the categorization is carried out by the categorizing means;

response category determining means for determining a response category information on which is stored in advance in correspondence to the category of the input voice the category of which is determined by the categorizing means;

feature determining means for determining voice feature information of the output response statement in correspondence to the response category determined by the response category determining means, the number of times counted by the categorization counting means, and the time information on which is stored in the time storing means; and

voice synthesizing means for obtaining a voice of the output response statement through synthesis based on the voice feature information determined by the feature determining means and the rhythm or spectrum determined by the rhythm/spectrum determining means.

[Detailed Description of the Invention]

[0001]

[Technical Field to which the Invention belongs] The present invention relates to a voice interactive method and a voice interactive device for smoothly carrying out information transmission between a human and a computer.

[0002]

[Prior Art] A voice interactive device for carrying out control while a user interacts with a computer using his/her voice is adopted in a car navigation system or a game machine, and its application is increasing.

[0003] A voice interactive device disclosed in JP 07-210193 A is known as a conventional voice interactive device. However, this conventional voice interactive device involves a problem that since an outputted synthetic voice is monotonous and thus voice quality, a speed, a tone, intonation, or the like is not changed, a user loses interest in a machine operation. In addition, JP 06-110650 A discloses a voice interactive device in which information on a part of an input voice is stored, and when a voice is outputted, a speed of the voice the information on which is stored is changed. However, neither of those voice interactive devices outputs a response voice in consideration of a feeling of a speaker, a progress of an interaction, or the like, and is said as being user-friendly. For example, in a case where a user uses a voice interactive device built in a car navigation system, when he/she cannot successfully find out his/her destination, he/she gradually becomes irritated, or when he/she is late for appointed time due to traffic congestion, he/she is discouraged. Thus, though a feeling of the speaker changes in correspondence to a situation, the conventional voice interactive device always responds just in the even tone.

[0004] In addition, there is encountered a problem that since the conventional output voice has no modulation and thus is outputted in a fixed tone, it is difficult to grasp an important point. For example, it is supposed that when a user retrieves a restaurant during traveling, firstly, the voice interactive device makes a response of "A sushi bar A exists in a position 3 km from here on the left-hand side" to the retrieval. Then, when a user voice-inputs "Any other?", the conventional car navigation system makes a response of "A family restaurant B exists in a position 5 km from here on the right-hand side" in the monotonous voice quality, speed, tone, intonation, and the like to the input voice. In this case, the important information is "The family restaurant B" corresponding to "Any other?". However, since the car navigation system makes a response in an even tone with respect to other information ("5 km", "from here", "on the right-hand side", and "exists") as well, there is encountered a problem that it is hard to grasp the important information. In particular, when the contents of a conversation between the voice interactive device and a human become long, the outputted information becomes complicated accordingly, and thus it may be said that such a bad effect becomes significant.

[0005]

[Problems to be solved by the Invention] The present invention has been made in light of such circumstances, and it is, therefore, an object of the present invention to provide a user-friendly voice

interactive method and voice interactive device which are capable of making a response after classification of an important point and recognition of a feeling of a speaker during a conversation.

[0006] It is another object of the present invention to provide a user-friendly voice interactive method and voice interactive device which are capable of making a response after flexibly coping with a feeling of a speaker which rapidly changes in correspondence to a progress of a conversation.

[0007] It is still another object of the present invention to provide a user-friendly voice interactive method and voice interactive device by which a user is easy to recognize an important point in an outputted response statement.

[0008]

[Means for solving the Problem] According to a first aspect of the present invention, in a voice interactive method of making a response statement in response to an inputted voice to output a voice, the voice interactive method is characterized by including: a step of converting an input voice into a voice signal; a step of converting the voice signal into voice feature information; a step of determining a category of the input voice from a predetermined category group based on the voice feature information; a step of determining a predetermined response category in correspondence to the category of the input voice; a step of determining voice feature information of an output response statement in correspondence to the response

category; and a step of obtaining a voice of the output response statement through synthesis based on the voice feature information determined in the feature information determining step.

[0009] According to a second aspect of the present invention, the voice interactive method according to the first aspect of the invention is characterized by further including a step of counting the number of times on a category basis by which the categories are determined, in which the feature information in the feature information determining step is determined in correspondence to the response category and the number of times counted in the number-of-times counting step.

[0010] According to a third aspect of the present invention, the voice interactive method according to the second aspect of the invention is characterized by further including a step of storing information on time on a category basis when the categories are determined, in which the feature information in the feature information determining step is determined in correspondence to the number of times counted in the number-of-times counting step and the information on the times stored in the time storing step.

[0011] According to a fourth aspect of the present invention, in the voice interactive method of making a response statement in response to an inputted voice to output a voice, the voice interactive method is characterized by including: a step of converting an input voice into a voice signal; a step of converting the voice signal

into voice feature information; a step of determining a keyword and a phoneme of the input voice from a predetermined keyword and phoneme group based on the voice feature information; a step of counting the number of times by which the same keyword is inputted; a step of making a response statement based on the inputted keyword and phoneme; a step of, with respect to response words of the made response statement which are made in correspondence to the keyword, determining a rhythm or a spectrum of each of the response words in the response statement to be outputted in correspondence to the number of times of keyword inputs; a step of determining a category of the input voice from a predetermined category group based on the voice feature information; a step of counting the number of times on a category basis by which the categories are determined; a step of storing information on time on a category basis when the categories are determined; a step of determining a predetermined response category in correspondence to the category of the input voice; a step of determining voice feature information of the output response statement in correspondence to the response category, the number of times counted in the category determining step, and the times the information on which is stored in the time storing step; and a step of obtaining a voice for the output response statement through synthesis based on the determined voice feature information of the output response statement, and the determined rhythms or spectra of the response words in the output response statement.

[0012] According to a fifth aspect of the present invention, in the voice interactive method of making a response statement in response to an inputted voice to output a voice, the voice interactive method is characterized by including: a step of converting an input voice into a voice signal; a step of converting the voice signal into voice feature information; a step of determining a keyword and a phoneme of the input voice from a predetermined keyword and phoneme group based on the voice feature information; a step of counting the number of times by which the same keyword is inputted; a step of making a response statement based on the inputted keyword and phoneme; a step of, with respect to response words of the made response statement which are made in correspondence to the keyword, determining a rhythm or a spectrum of each of the response words in the response statement to be outputted in correspondence to the number of times of keyword inputs; a step of determining a category of the input voice from a predetermined category group based on the voice feature information; a step of counting the number of times on a category basis by which the categories are determined; a step of storing information on time on a category basis when the categories are determined; a step of determining a predetermined response category in correspondence to the category of the input voice; a step of determining voice feature information of the output response statement in correspondence to the response category, the number of times counted in the category determining step, and the

times the information on which is stored in the time storing step; and a step of obtaining a voice for the output response statement through synthesis based on the determined voice feature information of the output response statement, and the determined rhythms or spectra of the response words in the output response statement.

[0013] According to a sixth aspect of the present invention, in the voice interactive device for making a response statement in response to an inputted voice to output a voice, the voice interactive device is characterized by including: voice converting means for converting an inputted voice signal into voice feature information; categorizing means for determining a category of the input voice from a category group information on which is stored in advance based on the voice feature information outputted from the voice converting means; response category determining means for determining a response category information on which is stored in advance in correspondence to the category of the input voice categorized by the categorizing means; feature determining means for determining voice feature information of an output response statement in correspondence to the response category determined by the response category determining means; and voice synthesizing means for obtaining a voice for the output response statement through synthesis based on the voice feature information determined by the feature determining means.

[0014] According to a seventh aspect of the present invention, the

voice interactive device according to the sixth aspect of the invention is characterized by further including categorization counting means for counting the number of times by which the categorization is carried out by the categorizing means on a category basis, in which the feature determining means determines voice feature information of an output response statement in correspondence to the response category determined by the response category determining means and the number of times counted by the categorization counting means.

[0015] According to an eighth aspect of the present invention, the voice interactive device according to the seventh aspect of the invention is characterized by further including time storing means for storing information on time when the categorization is carried out by the categorizing means on a category basis, in which the feature determining means determines voice feature information of an output response statement in correspondence to the response category determined by the response category determining means, the number of times counted by the categorization counting means, and the times the information on which is stored in the time storing means.

[0016] According to a ninth aspect of the present invention, in the voice interactive device for making a response statement in response to an inputted voice to output a voice, the voice interactive device is characterized by including: voice converting means for

converting an inputted voice signal into voice feature information; keyword determining means for determining a keyword and a phoneme of an input voice from a keyword and phoneme group information on which is stored in advance based on the voice feature information outputted from the voice converting means; keyword counting means for counting the number of times by which the same keyword is inputted in the keyword determining means; response statement making means for making a response statement based on the inputted keyword and phoneme; rhythm/spectrum determining means for, with respect to response words, of the response statement made by the response statement making means, which is made in correspondence to the keyword, determining a rhythm or a spectrum of each of response words in the response statement outputted in correspondence to the number of times of keyword inputs; categorizing means for determining a category of the input voice from a category group information on which is stored in advance based on the voice feature information outputted from the voice converting means; categorization counting means for counting the number of times by which the categorization is carried out by the categorizing means on a category basis; time storing means for storing information on time when the categorization is carried out by the categorizing means on a category basis; response category determining means for determining a response category information on which is stored in advance in correspondence to the category of the input voice the category of which is determined

by the categorizing means; feature determining means for determining voice feature information of the output response statement in correspondence to the response category determined by the response category determining means, the number of times counted by the categorization counting means, and the times information on which is stored in the time storing means; and voice synthesizing means for obtaining a voice of the output response statement through synthesis based on the voice feature information determined by the feature determining means and the rhythm or spectrum determined by the rhythm/spectrum determining means.

[0017] According to a tenth aspect of the present invention, in the voice interactive device for making a response statement in response to an inputted voice to output a voice, the voice interactive device is characterized by including: voice converting means for converting an inputted voice signal into voice feature information; keyword determining means for determining a keyword and a phoneme of an input voice from a keyword and phoneme group information on which is stored in advance based on the voice feature information outputted from the voice converting means; keyword counting means for counting the number of times by which the same keyword is inputted in the keyword determining means; response statement making means for making a response statement based on the inputted keyword and phoneme; rhythm/spectrum determining means for, with respect to response words, of the response statement made by the response

statement making means, which is made in correspondence to the keyword, determining a rhythm or a spectrum of each of response words in the response statement outputted in correspondence to the number of times of keyword inputs; categorizing means for determining a category of the input voice from a category group information on which is stored in advance based on the voice feature information outputted from the voice converting means; categorization counting means for counting the number of times by which the categorization is carried out by the categorizing means on a category basis; time storing means for storing information on time when the categorization is carried out by the categorizing means on a category basis; response category determining means for determining a response category information on which is stored in advance in correspondence to the category of the input voice the category of which is determined by the categorizing means; feature determining means for determining voice feature information of the output response statement in correspondence to the response category determined by the response category determining means, the number of times counted by the categorization counting means, and the times information on which is stored in the time storing means; and voice synthesizing means for obtaining a voice of the output response statement through synthesis based on the voice feature information determined by the feature determining means and the rhythm or spectrum determined by the rhythm/spectrum determining means.

[0018] In the first and sixth aspects of the present invention, the voice feature information such as a speed, intonation, a power, or a spectrum of the input voice is extracted. Then, the voice feature information of the input voice is compared with the voice feature information group stored on a category basis such as "doubt", "discouragement", or "irritation". As a result of the comparison, the category corresponding to the most approximate voice feature information is determined as the category of the input voice. Moreover, the information on the response category group such as "kind response", "speedy response", or "polite response" is stored in advance, and a suitable response category is determined in correspondence to the category of the input voice. Also, the voice feature information such as the voice speed, the intonation, the power, and the spectrum is synthesized in correspondence to the kind of response category. Hence, for example, it becomes possible that if a speaker rapidly put a question to the voice interactive device, the voice interactive device makes a response at a high speed to the speaker. Thus, it becomes possible to construct a user-friendly voice interactive system which is free from the monotony and thus in which a user has a feeling as if he/she talks with a human.

[0019] In the second and seventh aspects of the present invention, the number of times by which the categorization is carried out is counted, and the voice feature information such as the voice speed,

the intonation, the power, and the spectrum of the output response statement is synthesized in correspondence to the number of times. Hence, it becomes possible to construct a user-friendly voice interactive system which is freer from the monotony and in which a feeling of a speaker changing from moment to moment during the interaction is also taken into consideration.

[0020] In the third and eighth aspects of the present invention, the information on the number of times by which the categorization is carried out and the information on the time when the categorization is carried out are stored, and the voice feature information such as the voice speed, the intonation, the power, and the spectrum of the output response statement is synthesized in correspondence to the number of times and the time of the categorization. Hence, it becomes possible to construct a more user-friendly voice interactive system which is much freer from the monotony and in which a change of mind of a speaker is taken into consideration in terms of a long term.

[0021] In the fourth and ninth aspects of the present invention, the number of times by which the same keyword is inputted is counted. Then, the rhythm or the spectrum is changed for the response word corresponding to the keyword having a high input frequency. For example, the response word corresponding to the keyword having a high input frequency is outputted with a larger sound volume than that corresponding to the keyword having the low input frequency.

Hence, a speaker can exactly catch the important point in the response voice.

[0022] In the fifth and tenth aspects of the present invention, the number of times by which the same keyword is inputted is counted. In addition, the response category for the input voice category is determined, and the number of times by which the categorization is carried out is counted and the information on the time when the categorization is carried out is stored. Also, the voice feature information in the output response statement and the rhythm or the spectrum of the response words in the output response statement are synthesized based on the response category, the number of times by which the categorization is carried out, the time when the categorization is carried out, and the number of times by which the keyword is inputted. Hence, it is possible to construct a user-friendly voice interactive system which is free from the monotony and in which a user is easy to catch an important point in an output voice.

[0023]

[Embodiment Mode of the Invention] The present invention will hereinafter be described in detail based on the drawings showing an embodiment mode thereof. Fig. 1 is a block diagram showing a voice interactive device D of the present invention. In Fig. 1, reference numeral 1 designates a microphone for converting an input voice into an electrical voice signal. The electrical voice signal

is digitized by an A/D converter 4. The digitized voice signal is then converted into voice feature information by an MPU 2. In addition, for the digitized voice signal, the MPU 2 executes a processing for detecting a keyword or a phoneme, a processing for making a response statement, a processing for determining a category, and the like. Also, the MPU 2 includes in its inside a clock portion 7. Note that the processings executed by the MPU 2 will be described later.

[0024] In addition, a storage device 3 such as a hard disk is connected to the MPU 2. Various kinds of information such as keyword data, phoneme data, and category data are stored in the storage device 3. A control signal and various kinds of information such as distance information data or a traffic congestion information text file are transmitted between the voice interactive device D, and a car navigation system (not shown), a game apparatus (not shown), or the like through an I/O portion 8. Data such as response statement data, sound volume data, and voice feature information generated in the MPU 2 is converted into an electrical signal by a D/A converter 5. The resultant electrical signal is then outputted in the form of a voice through a speaker 6.

[0025] Fig. 2 is a flow chart showing a processing procedure in the MPU 2. Firstly, the digital voice signal outputted from the A/D converter 4 is converted into voice feature information such as a speed, intonation, a power, or a frequency spectrum obtained

through FFT (Step S21). A keyword processing for detecting a keyword and the like to make a response statement and the like, and a category processing for detecting a feeling category of a speaker are executed for the voice feature information processed in Step S21.

[0026] Firstly, the category processing will be described hereinafter. The voice feature information (such as a speed, intonation or a frequency spectrum of an input voice) is stored on a category basis (such as a doubt, confirmation or irritation) in a category database 3d in the storage device 3. Firstly, a category of an input voice is determined based on the voice feature information obtained through the conversion in Step S21 (Step S22). When a category of the input voice is determined, the voice feature information of the input voice is compared with the voice feature information stored in the category database 3d, and a category corresponding to the voice feature information which agrees with or is most approximate to the voice feature information of the input voice is determined as the category of the input voice (speaker) (Step S22).

[0027] Then, a counter is provided on a category basis determined through the categorization in Step S22. Whenever the same category is inputted, the contents of the corresponding counter are incremented (Step S23). In addition, information on the times when the categorization is carried out in Step S23 is successively stored on a category basis (Step S24). The information on the times is

stored in accordance with a signal from the clock portion 7, and is successively stored whenever the categorization is carried out in Step S22. Information on a response category group corresponding to the category of the input voice is stored in advance in a response category database 3e in the storage device 3. A response category corresponding to the category of the input voice obtained through the categorization in Step S22 is determined in consideration of the response category database 3e (Step S25). Voice feature information corresponding to the response category is stored in a tune database 3f of the storage device 3. The voice feature information corresponding to the response category determined in Step S25 is determined by referring to the tune database 3f (Step S26).

[0028] Next, the keyword processing will be described. Voice feature information is stored on a keyword basis in a keyword database 3a of the storage device 3. Likewise, voice feature information is stored on a phoneme basis in a phoneme database 3b of the storage device 3. Firstly, the voice feature information processed in Step S21 is compared with the voice feature information stored in the keyword database 3a and the phoneme database 3b. Then, a keyword and a phoneme corresponding to the voice feature information which agrees with the voice feature information processed in Step S21 are determined (Step S27).

[0029] Subsequently, a counter is provided for every keyword in

Step S27. Whenever the same keyword is inputted, the contents of the corresponding counter are incremented (Step S28). Then, a response statement is made based on the keyword and the phoneme determined in Step S27 (Step S29). Pronunciation symbols and meaning contents of the keywords and the phonemes are stored in a dictionary database 3c of the storage device 3. In a response statement making processing (Step S29), a response statement is made while the MPU 2 refers the dictionary database 3c. Then, a rhythm or a spectrum is determined every response word corresponding to the keyword in the response statement made in Step S29 in correspondence to the number of times by which the keywords are counted in Step S28 (Step S210). Note that the rhythm means a height, a speed, strength, or the like of a voice.

[0030] When the above-mentioned procedure of the keyword processing and the category processing is completed, the results of those two processings are reflected in a voice for the response statement. In other words, the voice for the response statement made in Step S29 is obtained through synthesis based on the voice feature information determined in Step S26 and the rhythm or spectrum information of the response words corresponding to the keyword determined in Step S210 (Step S211). The processing contents will be described in detail hereinafter. Note that any of the keyword processing and the categorization processing may be firstly processed.

[0031] Fig. 3 is a diagram explaining a procedure of the categorization processing. In Fig. 3, the voice feature information is stored as predefined values on a category basis in the category database 3d. The voice feature information, for example, has a speed, intonation, a power, and a spectrum. For example, in a case of the "doubt" category, such information that the ending of a word rises is stored in the category database 3d.

[0032] In the categorization processing (Step S22), the voice feature information of the inputted voice is compared with the voice feature information which is stored as the predefined values on a category basis in the category database 3d by utilizing a method such as a pattern matching method. As a result of the comparison, the category the information on which is stored in the category database 3d in correspondence to the agreed or most approximate voice feature information is determined as the category of the input voice. Note that in order to enhance the precision of the category determination, a user may carry out dictation and the voice feature information stored in the category database 3d may be suitably updated based on the dictation results. Moreover, in a case where the voice interactive device according to the present invention is utilized by two or more persons, the dictation results for each user are classified to be stored.

[0033] Fig. 4 is a flow chart showing a categorization processing procedure. Firstly, the voice feature information of the input voice

obtained through the conversation in Step S21 is compared with the voice feature information which is stored as the predefined values on a category basis in the category database 3d (Step S41). Then, the optimal one voice feature information is determined from the voice feature group in the category database 3d (Step S42). Thereafter, the category corresponding to such voice feature information is determined as the category of the input voice (Step S43).

[0034] Fig. 5 is a diagram explaining the processing procedure for the response category determination. The information on the response categories corresponding to the input categories is stored in the response category database 3e. For example, when the input voice category is "haste", it is better that the output voice is outputted in a crisp manner, and hence the information on the response category corresponding to the input category is stored in the response category database 3e so that the response category become "speedy response". In addition, when the input voice category is "discouragement", since the output voice for calming the user is preferable, the information on an "encouraging response" is stored as the response category corresponding to the input category in the response category database 3e. In the response category determining processing (Step S25), there is executed a processing for determining the response category corresponding to the category determined in the categorization processing (Step S22) based on

the information stored in the response category database 3e.

[0035] The voice feature information of the response voices is stored on a response category basis in the tune database 3f. For example, in a case of a "speedy response", the predefined speed "C (slightly speedy speed)" as one of the voice feature information is selected, and moreover the predefined intonation "ノ(fixed intonation)" and the predefined power "c" are selected. In the voice feature information, the predefined values may be changed based on the setting by a user. In the feature determining processing (Step S26), there is executed a processing for determining the voice feature information of the voice which is outputted based on the response category determined by the response category determining processing (Step S25) by referring the tune database 3f.

[0036] Fig. 6 is a flow chart explaining a procedure of the categorization counting processing and the feature determining processing. In the categorization counting processing (Step S23), there is executed a processing for storing the information on the number of times on a category basis by which the categorization is carried out in the categorization processing (Step S22). In a case of an example of Fig. 6, the information on the "doubt" category, the information on the "irritation" category, and the information on the "haste" are stored "one" time, "three" times, and "two" times, respectively. In the response category determining processing (Step S25), the response category corresponding to the input category

is determined, and likewise, the number of times of the response category to the number of times of the input categorization is also counted. In the case of the example of Fig. 6, the response category of a "kind response" corresponding to the input category of the "doubt", the response category of a "polite response" corresponding to the input category of "irritation", and the response category of a "speedy response" corresponding to the input category of "haste" are counted "one" time, "three" times, and "two" times, respectively.

[0037] In the feature determining processing (Step S26), there is executed a processing for determining the voice feature information in consideration of the response category and the number of times determined in Step S25. For example, when the "kind response" is continuously made, the control is carried out so that the second response becomes the "kinder response" than that in the first response. When the "polite response" continues, the voice feature of the response voice is controlled so as to provide the "politer response". Even when the same response category does not continue, the control is flexibly carried out in correspondence to the counted number of times.

[0038] Fig. 7 is a flow chart explaining a procedure of the time storing processing. In the time storing processing (Step S24), there is executed a processing for successively storing information on time on a category basis when the categorization is carried out in the categorization processing (Step S22). Then, the voice feature

information is determined in consideration of the time the information on which is stored. For example, it is supposed that after the category is determined as "irritation", the category of the voice which is inputted next time is "discouragement". In this case, the category of a "polite and encouraging response" or the like is selected as the response category in consideration that the category of the input voice changes from "irritation" to "discouragement" in a time series manner. In response thereto, in the feature detecting processing (Step S26), the voice feature information is determined by utilizing the same method as that in the above-mentioned categorization counting processing (Step S23). Thus, the response voice is outputted in which a long term factor is also taken into consideration.

[0039] The counted number of times in the categorization counting processing (Step S23) and the time the information on which is stored through the time storing processing (Step S24) are updated whenever necessary until a reset request is made from a control portion (not shown) of the voice interactive device. Since in the car navigation system for example, the voice interaction is continuously carried out until one task is completed (e.g., the retrieval is completed when a destination is retrieved), the number of times and the time are not cleared for this period of time. Conversely, when an operation proceeds to another task (e.g., when an operation proceeds to guidance for traffic congestion information after the retrieval

for a destination is completed), the number of times and the time are cleared to be initialized since the interaction contents are different from those in the former task. When a request for updating the number of times and the time to "0" is made in such a manner, the number of times and the time are cleared to "0".

[0040] Fig. 8 is a flow chart showing a procedure of a processing for updating the number of times of the categorization and the time of the categorization. Firstly, the category is determined, the number of times by which the category is counted, and the information on the time when the categorization is carried out is stored (Step S81). Thus, the response category is determined based on the input voice category (Step S82). Next, the voice feature information is determined based on the determined response category, the number of times of the categorization, and the time of the categorization (Step S83). Then, when no request for clearing the number of times and the time is made from the control portion (not shown) (No in Step S84), an operation returns back to Step S81.

[0041] Then, when a voice is inputted again to be categorized, the number of times which is counted in correspondence to the category is incremented (Step S81). At the same time, the information on the time when the categorization is carried out is successively and additionally stored on a category basis (Step S81). At this time, the weighting is carried out in correspondence to the counted value and the stored information on the time, and then the voice

feature information is determined (Step S83). On the other hand, when the request for clearing the number of times and the time is made from the control portion (not shown) (Yes in Step S84), the number of times of the categorization is cleared to "0" and the information on the time stored is all erased (Step S85).

[0042] Fig. 9 is a flow chart explaining a procedure of the keyword processing of the present invention. In the keyword determining processing (Step S27), there is executed a processing for counting every keyword whenever the keyword is detected (Step S28). For example, when a user retrieves a convenience store using the car navigation system, the user utters a sound of "convenience store", the keyword, "convenience store", is counted as "1". Note that the number of times of the keyword counting is also accumulated similarly to the case of the number of times of the categorization until a clear request is made from the control portion (not shown).

[0043] In the response statement making processing (Step S29), there is executed a processing for accessing the dictionary database 3c and a map database (not shown) of the car navigation system to retrieve a convenience store based on the keyword in order to make a response statement. Then, a response statement of "Hyson is located at 3 km from here on the right-hand side" is made. In the rhythm/spectrum determining processing (Step S210), a response word, "Hyson", corresponding to the keyword, "convenience store", in the response statement is weighted as "1" based on the number of times of the

keyword counting. On the other hand, since with respect to other response words, "at 3 km from here on the right-hand side" and "is located", the number of times of the keyword counting is "0", a processing for weighting other response words as "0" is executed. Then, a processing for determining a rhythm or a spectrum based on the weight is executed. For example, the sound volume for the response words of "at 3 km from here on the right-hand side" and "exists" is determined as the normal sound volume, and the sound volume for the response word of "Hyson" is determined as the higher sound volume than the normal sound volume.

[0044] When the interaction further continues, the number of inputted keywords is further counted. To be specific, when "any other?" is subsequently inputted, the keyword "any other?" is counted as "1". Then, when the response statement is "any other convenience store is not located around here", since each of the keywords "convenience store" and "any other?" is counted in number of times as "1", the response word "any other", corresponding to the keyword "any other?", and the response word "is not located", corresponding to the keyword "convenience store" are weighted with the same value, and thus the response words "any other?" and "is not located" are outputted with the same sound volume. On the other hand, since other response words, "around here" is "0" in number of times of the counting, i.e., "0" in weight, the sound volume of other response words, "around here" is set lower than that in "any other?" and "is not located".

[0045] It is supposed that the user further produces a sound of "any other? where?". Then, the keywords, "convenience store", "any other?", and "where?" are counted "one" time, "two" times, and "one" time, respectively. In this case, it is supposed that the response statement is "Seven is located as a convenience store other than Hyson at 10 km from here on the left-hand side. Any other convenience store other than Seven is at a distance of more than 10 km from here". Then, the response word "other than" of the keyword "any other?", the response words "Hyson" and "Seven" of the keyword "convenience store", the response words "at 10 km from here on the left-hand side" and " at a distance of more than 10 km from here" of the keyword "where?", and the response words "on", "is located. ... any other convenience store other than Seven", and "is" other than those response words, are weighted in descending order of the detection number of times as "2", "1", "1", and "0" in correspondence to the numbers of times of the counting, respectively. As a result, the sound volumes are determined in descending order of the weight based on such weights.

[0046] Fig. 10 is a flow chart explaining a procedure of the voice synthesis processing. In the above-mentioned example, the conventional voice interactive device outputs the response statement of "Seven is located as a convenience store other than Hyson at 10 km from here on the left-hand side. Any other convenience store other than Seven is at a distance of more than 10 km from here",

which is evenly made without the emphasis, the intonation, the feeling, and the like. However, the present invention executes the processing for determining the rhythm or spectrum on a response word basis based on the above-mentioned weights obtained through the rhythm/spectrum determining processing (Step S210). On the other hand, the voice feature information corresponding to the response category ("kind response" in the example in Fig. 10) determined through the response category determining processing (Step S25) is determined through the feature determining processing (Step S26) (in the example of Fig. 10, the speed of "slow" and the intonation of "the ending of a word is lowered" are determined). Thus, there is executed a processing for adding the voice feature information, and the rhythm or spectrum information to the voice data of the response statement made through the response statement making processing (Step S29) based on the rhythm or spectrum determined in the rhythm/spectrum determining processing (Step S210) or the voice feature information determined through the feature determining processing (Step S26) (the voice synthesizing processing in Step S211). As a result, the response voice in which a user is easy to catch to an important point is outputted in a feeling-like manner.

[0047]

[Effects of the Invention] As set forth hereinabove in detail, in the first and sixth aspects of the present invention, the voice feature information such as a speed, intonation, a power, and a

spectrum of the input voice is extracted. Then, the voice feature information of the input voice is compared with the voice feature information group which are stored on a category basis such as "doubt", "discouragement", or "irritation". As a result of the comparison, the category corresponding to the most approximate voice feature information is determined as the category of the input voice. Moreover, the information on the response category group including a "kind response", a "speedy response, or a "polite response" is stored in advance, and one suitable response category is determined in correspondence to the category of the input voice. Then, the voice feature information, such as the voice speed, the intonation, the power, and the spectrum is synthesized in correspondence to a kind of response category. Here, for example, it becomes possible that if a speaker asks the voice interactive device a question in haste, the voice interactive device makes a response to the speaker at a high speed. As a result, it becomes possible to construct the user-friendly voice interactive system which is free from the monotony and in which a user feels like as if he/she has a conversation with a human.

[0048] In addition, in the second and seventh aspects of the present invention, the number of times by which the categorization is carried out is counted, and the voice feature information such as the voice speed, the intonation, the power, and the spectrum in the output response statement is synthesized in correspondence to the number

of times of the categorization. Hence, it becomes possible to construct the user-friendly voice interactive system which is freer from the monotony and in which a feeling of a speaker which changes from moment to moment is taken into consideration.

[0049] In addition, in the third and eighth aspects of the present invention, the information on the number of times by which the categorization is carried out and the information on the time of the categorization are stored, and the voice feature information such as the voice speed, the intonation, the power, and the spectrum in the output response statement is synthesized in correspondence to the number of times of the categorization and the time of the categorization. Hence, it becomes possible to construct the user-friendly voice interactive system which is much freer from the monotony and in which a change of mind of a speaker is taken into consideration in terms of a long term.

[0050] In addition, in the fourth and ninth aspects of the present invention, the number of times by which the same keyword is inputted is counted. Then, the rhythm or spectrum of the response word corresponding to the keyword having a high input frequency is changed to output that response word, e.g., the response word corresponding to the keyword having a high input frequency is outputted with a larger sound volume than that corresponding to the keyword having a low input frequency. Hence, a speaker can exactly catch an important point in the response voice.

[0051] Moreover, in the fifth and tenth aspects of the present invention, the number of times by which the same keyword is inputted is counted. In addition, the response category corresponding to the input voice category is determined, the number of times by which the categorization is carried out, and the information on the time when the categorization is carried out is stored. Then, the voice feature information in the output response statement, and the rhythm or spectrum of each of the response words in the output response statement are synthesized based on the response category, the number of times by which the categorization is carried out, the time when the categorization is carried out, and the number of times of the keyword inputs. Hence, it becomes possible to construct the user-friendly voice interactive system which is free from the monotony and in which a user is easy to catch an important point of the output voice.

[Brief Description of the Drawings]

[Fig. 1] A block diagram showing a voice interactive device of the present invention.

[Fig. 2] A flow chart showing a processing procedure in an MPU.

[Fig. 3] An explanatory diagram showing a procedure of a categorizing processing.

[Fig. 4] A flow chart showing a categorizing processing procedure.

[Fig. 5] An explanatory diagram showing a processing procedure of response category determination.

[Fig. 6] An explanatory diagram showing a procedure of a categorization counting processing and a tune determining processing.

[Fig. 7] An explanatory diagram showing a procedure of a time storing processing.

[Fig. 8] A flow chart showing a procedure of a processing for updating the number of times of categorization and time of the categorization.

[Fig. 9] An explanatory diagram showing a procedure of a keyword processing of the present invention.

[Fig. 10] An explanatory diagram showing a procedure of a voice synthesizing processing.

[Description of Reference Symbols]

D voice interactive device

1 microphone

2 MPU

3 storage device

3a keyword database

3b phoneme database

3c dictionary database

3d category database

3e response category database

3f tune database

6 speaker

FIG. 1

3 STORAGE DEVICE

4 A/D CONVERTER

5 D/A CONVERTER

7 CLOCK PORTION

8 I/O PORTION

FIG. 2

3a KEYWORD DATABASE

3b PHONEME DATABASE

3c DICTIONARY DATABASE

3d CATEGORY DATABASE

3e RESPONSE CATEGORY DATABASE

3f TUNE DATABASE

START

S21 VOICE CONVERTING PROCESSING

CATEGORY PROCESSING

S22 CATEGORIZING PROCESSING

S23 CATEGORIZATION COUNTING PROCESSING

S24 TIME STORING PROCESSING

S25 RESPONSE CATEGORY DETERMINING PROCESSING

S26 FEATURE DETERMINING PROCESSING

KEYWORD PROCESSING

S27 KEYWORD DETERMINING PROCESSING

S28 KEYWORD COUNTING PROCESSING

S29 RESPONSE STATEMENT MAKING PROCESSING

S210 RHYTHM/SPECTRUM DETERMINING PROCESSING

S211 VOICE SYNTHESIS

FIG. 3

3d CATEGORY DATABASE

CATEGORY

DOUBT

CONFIRMATION

VOICE FEATURE INFORMATION

SPEED

INTONATION

POWER/SPECTRUM

VOICE FEATURE INFORMATION

S22 CATEGORIZING PROCESSING

FIG. 4

S41 COMPARE

S42 DETERMINE VOICE FEATURE INFORMATION

S43 DETERMINE CATEGORY

FIG. 5

3e RESPONSE CATEGORY DATABASE

CATEGORY

DOUBT

CONFIRMATION

HASTE

DISCOURAGEMENT

IRRITATION

RESPONSE CATEGORY

KIND RESPONSE

SLOW RESPONSE

SPEEDY RESPONSE

ENCOURAGING RESPONSE

POLITE RESPONSE

S22 CATEGORIZING PROCESSING

S25 RESPONSE CATEGORY DETERMINING PROCESSING

S26 FEATURE DETERMINING PROCESSING

3f TUNE DATABASE

SPEED

INTONATION

POWER

FIG. 6

S22 CATEGORIZING PROCESSING

S23 CATEGORIZATION COUNTING PROCESSING

DOUBT

IRRITATION

HASTE

S25 RESPONSE CATEGORY DETERMINING PROCESSING

KIND RESPONSE

POLITE RESPONSE

SPEEDY RESPONSE

S26 FEATURE DETERMINING PROCESSING

FIG. 7

S23 CATEGORIZATION COUNTING PROCESSING

IRRITATION

DISCOURAGEMENT

S24 TIME STORING PROCESSING

HOUR, MINUTE, SECOND

S25 RESPONSE CATEGORY DETERMINING PROCESSING

POLITE AND ENCOURAGING RESPONSE

S26 FEATURE DETERMINING PROCESSING

FIG. 8

START

S81 COUNT NUMBER OF TIMES OF CATEGORY DETERMINATION AND STORE TIMES

S82 DETERMINE RESPONSE CATEGORY

S83 DETERMINE FEATURE

S84 ARE NUMBER OF TIMES AND TIME CLEARED?

S85 CLEAR NUMBER OF TIMES AND TIME

END

FIG. 9

S27 KEYWORD DETERMINING PROCESSING

S28 KEYWORD COUNTING PROCESSING

"CONVENIENCE STORE", "ANY OTHER?", "WHERE?"

S29 RESPONSE STATEMENT MAKING PROCESSING

HYSON IS LOCATED AT 3 km FROM HERE ON RIGHT-HAND SIDE

S210 RHYTHM/SPECTRUM DETERMINING PROCESSING

"CONVENIENCE STORE" → "HYSON" → LARGE

"... AT 3 km FROM HERE ON RIGHT-HAND SIDE" → MIDDLE

"... IS LOCATED" ... → MIDDLE

EXAMPLE OF CONVERSATION

D: VOICE INTERACTIVE DEVICE

U: USER

U: CONVENIENCE STORE

D: HYSON IS LOCATED AT POSITION 3 km FROM HERE.

U: ANY OTHER?

D: ANY OTHER CONVENIENCE STORE IS NOT LOCATED AROUND HERE.

U: ANT OTHER? WHERE?

D: SEVEN IS LOCATED AS CONVENIENCE STORE OTHER THAN HYSON AT 10 km FROM HERE ON LEFT-HAND SIDE. ANY OTHER CONVENIENCE STORE OTHER THAN SEVEN IS AT DISTANCE OF MORE THAN 10 km FROM HERE.

FIG. 10

S25 RESPONSE CATEGORY DETERMINING PROCESSING

POLITE RESPONSE

S26 FEATURE DETERMINING PROCESSING

KIND RESPONSE → SPEED "SLOW", INTONATION "ENDING OF WORD IS LOWERED"

S29 RESPONSE STATEMENT MAKING PROCESSING

RESPONSE STATEMENT

"SEVEN IS LOCATED AS CONVENIENCE STORE OTHER THAN HYSON AT 10 km FROM HERE ON LEFT-HAND SIDE. ANY OTHER CONVENIENCE STORE OTHER THAN SEVEN IS AT DISTANCE OF MORE THAN 10 km FROM HERE."

S210 RHYTHM/SPECTRUM DETERMINING PROCESSING

"OTHER THAN" ... LARGE

"HYSON" "SEVEN" ... MIDDLE

"AT 10 km FROM HERE ON LEFT-HAND SIDE" "AT DISTANCE OF MORE THAN 10 km FROM HERE" ... MIDDLE

"ON" "IS LOCATED. ... ANY OTHER CONVENIENCE STORE OTHER THAN ..." ... SMALL

"IS ..." ... SMALL

S211 VOICE SYNTHESIS

SPEED "SLOW"

END OF WORD "LOWERED"

"SEVEN (MIDDLE) IS LOCATED AS CONVENIENCE STORE (SMALL) OTHER THAN (LARGE) HYSON (MIDDLE) AT 10 km FROM HERE ON LEFT-HAND SIDE (MIDDLE).

ANY OTHER CONVENIENCE STORE OTHER THAN (SMALL) SEVEN (MIDDLE) IS (SMALL) AT DISTANCE OF MORE THAN 10 km FROM HERE (MIDDLE)."

PATENT ABSTRACTS OF JAPAN

(11)Publication number : **2001-272991**
(43)Date of publication of application : **05.10.2001**

(51)Int.CI. G10L 13/06
G10L 13/00
G10L 13/08
G10L 15/00
G10L 15/28
G10L 15/22

(21)Application number : **2000-084948**
(22)Date of filing : **24.03.2000**

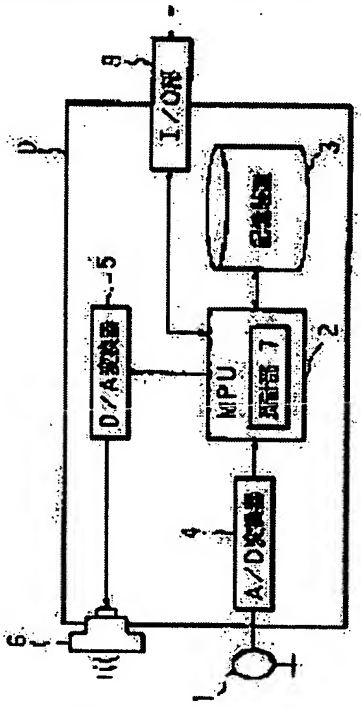
(71)Applicant : **SANYO ELECTRIC CO LTD**
(72)Inventor : **HASHIMOTO MAKOTO**

(54) VOICE INTERACTING METHOD AND VOICE INTERACTING DEVICE

(57)Abstract:

PROBLEM TO BE SOLVED: To solve the problem that equipment with a voice interactive function such as a car navigation device and a TV game, apears, however, such equipment does not exist as varies voice quality and tone of output voices according to progress situation of voice interactions, but a synthesized voice is outputted in the same voice quality and tone, therefore, a user loses interest in the machine operation.

SOLUTION: Since feelings of a speaker are categorized by analyzing speed or intonation or the like of input voices and responding voices are also categorized correspondingly to the categories so as to respond with feelings, the monotonousness is eliminated, and it becomes possible to construct a user-friendly voice interactive system like a conversation with a human being. Further, responding voices are outputted with more emphasis on responding words corresponding to higher frequent input keywords than those corresponding to lower frequent input words, therefore, the speaker can exactly hear important points among the responding voices.



LEGAL STATUS

[Date of request for examination] 21.04.2004

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]